DOCUMENT RESUME

ED 129 924 TM 005 801

AUTHOR Kosecoff, Jacqueline; Fink, Arlene
TITLE The Feasibility of Using Criterion-Referenced Tests for Large-Scale Evaluations.
PUB DATE Apr 76
NOTE 58p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Criteria; *Criterion Referenced Tests; Definitions; *Feasibility Studies; *Program Effectiveness; *Program Evaluation; Scores; Test Construction; Test Interpretation; Test Reliability; Test Reviews; Test Selection; Test Validity

ABSTRACT

The feasibility of using criterion referenced tests (CRTs) in a large-scale evaluation conducted in an effectiveness evaluation context was investigated. The study began by examining the theory that structures the development and validation of CRTs to discover whether, on theoretical grounds alone, CRTs are suitable or not suitable for large-scale effectiveness evaluations. Next, a set of criteria were developed for selecting tests appropriate for such evaluations. Included within the set of criteria was the stipulation that the test be able to provide scores amenable to CRT interpretation. Twenty-eight currently available CRTs were then reviewed, using the set of criteria. Finally, based on theoretical examination and the review, conclusions were drawn. Based on practical, not theoretical, considerations, it was concluded that there is no currently available CRT that is feasible for use in large-scale effectiveness evaluations. (RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED129924

The Feasibility of Using Criterion-Referenced Tests for Large-Scale Evaluations

Jacqueline Kosecoff, Ph.D. and Arlene Fink, Ph.D*
Center for the Study of Evaluation
University of California at Los Angeles

*The authors wish to express their gratitude to Penelope Morgan who contributed important ideas to this investigation and assisted in reviewing tests.

TM005 801

A paper presented at the Annual Meeting
of the American Educational Research Association.
San Fransisco, April 1976.

Criterion-referenced tests are becoming increasingly popular among educators and psychometricians. Perhaps the most important reason for their appearance and widespread acceptance can be traced to the new ways that had to be found to measure the effects of the educational reforms of the 1950's and 1960's. During those decades, the conventional school curriculum was declared in need of reform, and a reassessment of the goals and objectives of American education was made. (Hofstadter, 1963; Davis and Diamond, 1974; Cronbach and Suppes, 1969). Innovative courses of study and instructional technologies were subsequently developed, and programmed learning and individualized instruction became commonly-used teaching approaches. New ways of assessing student performance were needed that corresponded to the innovations.

Educators have traditionally relied on paper and pencil achievement tests to measure learning, so it was natural for them to turn to test theoreticians to provide them with alternative ways of interpreting performance on measures of educational achievement for the new curriculum and instruction. The psychometricians responded by pointing to two basic ways of assigning meaning to test scores. The first involved comparing one person's or group's performance or behavior with another person's or group's, and the second involved describing what a person or group can do or can be expected to do. Glaser (1963) referred to these two ways of giving meaning to test scores as "norm-referenced" and "criterion-referenced," and recommended criterion-referenced score interpretations for the reformed curriculum and instruction. According to Glaser and his colleagues, "A criterion-referenced test is one that is deliberately constructed to give scores that tell what kinds of behavior individuals with those score can demonstrate" (Glaser and Nitko, 1971).

The reaction to criterion-referenced tests (CRTs) was enthusiastic from the start. Because they provide score interpretations in terms of the achievement of specific and measurable skills and behaviors, CRTs have had appeal to those directly responsible for the education of students and the development and evaluation of educational programs. They also have had appeal to teachers who found the results of standardized tests inadequate to assist them in planning lessons, and to many educators and psychologists who judged standardized, norm-referenced tests to be unfair and even biased against individuals from under-privileged and minority groups. Finally, because the criterion-referenced approach was new, people saw it as an opportunity to improve on some of the mistakes they perceived to be built into norm-referenced testing.

CRT's popularity and sanction by theoreticians and practitioners has led to their frequent use for instructional diagnosis and placement and for measuring student achievement on educational tasks or objectives. In addition, CRTs are being suggested or used for other purposes like the evaluation of educational programs and the National Assessment of Educational Progress (Wilson, 1974). In fact, many recently-issued requests for proposals from state and federal agencies to evaluate educational programs have specifically required prospective contractors to justify their selection of standardized rather than CRT measures.

The purpose of this paper is to investigate the feasibility of using criterion-referenced tests in a large-scale evaluation conducted in an effectiveness evaluation context.

The investigation began by examining the theory that structures the development and validation of CRTs to discover whether, on theoretical grounds alone, CRTs are suitable or not suitable for large-scale effectiveness evaluations. The next step was to develop a set of criteria for selecting tests appropriate for such evaluations. Included within the set of criteria was the stipulation that the test be able to provide scores amenable to CRT interpretation. Currently available CRTs were then reviewed, using the set of criteria. Finally, based on the theoretical examination and the review, conclusions were drawn. This paper describes the investigation, and is organized into four parts:

. The Effectiveness Evaluation Context

. A Theoretical Examination of Criterion-Referenced Testing

. Review of Currently Available CRTs

. Conclusions

The Effectiveness Evaluation Context

Evaluation is a set of procedures used to appraise an educational program's merit and to provide information about the nature and quality of the program's goals, outcomes, impact, and costs (Fink and Kosecoff, 1976).

Evaluation Contexts

There are two contexts in which evaluations of educational programs are conducted. In one context, an evaluation is conducted to improve a program,

and the evaluation's clients are typically the program's organizers and staff. In the second context, an evaluation is conducted to measure the effectiveness of a program, and the evaluation's clients are typically the program's sponsors. The context for an evaluation is determined by the information needs of the individuals and agencies who must use the evaluation information.

An evaluation is performed in an improvement context when the evaluation's clients are concerned with finding out precisely where a change would make the program better. Typically, the organizers of a still-developing program require this kind of information so that they can modify and improve the program. On the other hand, an evaluation is conducted in an effectiveness context when the evaluation's clients are particularly concerned with determining the consistency and efficiency with which the program achieves desired results. Those individuals who sponsored program development, or who are interested in using the program, require this kind of information about a well-established program's outcomes and impact. In addition, in an effectiveness context, the evaluator usually makes use of powerful, experimental design strategies that permit comparisons, rely on empirically-validated and standardized instruments, and employ statistical and other analytic methods that allow inferences regarding the program's comparative value. Finally, in an effectiveness evaluation, the evaluator usually assumes a more global and independent stance toward the program than in an improvement context.

It is generally agreed (e.g., Alkin, et al, 1974) that information collection strategies for large-scale evaluations should rely upon instruments that have been demonstrated to be valid and reliable for the target population, and that are known to provide relevant information.

A Theoretical Examination of Criterion-Referenced Testing

In this section, theoretical issues in the development and validation of CRTs will be discussed. These include a definition of CRTs, the formulation and generation of CRT objectives and items, score interpretation schemes, establishing item and test quality, and the use of classical indexes of reliability and validity. Based on this discussion, the theoretical appropriateness of CRTs in effectiveness evaluation contexts will be investigated.

Definition

A criterion-referenced test is one that is designed to provide a measure of the extent to which educational purposes or tasks have been achieved. All CRTs share several features in common:

1. They are based on clearly-defined educational tasks and purposes.
2. Test items are specifically designed to measure the purposes and tasks.

3. Scores are interpreted in terms of attainment of a pre-set criterion or level of competence with respect to the purposes and tasks.

Other definitions of CRTs have also been offered. Three of the most often-used definitions are:

1. "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards...Performance standards are generally specified by defining a class or domain of tasks that should be performed by the individual" (Glaser and Nitko, 1971).

2. "A pure criterion-referenced test is one consisting of a sample of production tasks drawn from a well-defined population of performances, a sample that may be used to estimate the proportion of performances in that population at which the student can succeed" (Harris and Stewart, 1971).

3. "Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., a performance standard" (Popham and Husek, 1969).

While these definitions differ considerably in terms of the limitations and constraints placed on a criterion-referenced test, they all involve reporting test scores in terms of achievement of educational tasks.

A question frequently asked about criterion-referenced tests concerns their relationship to norm-referenced tests. To answer this question briefly, the crucial difference between these tests is the metric used to describe their scores. Norm-referenced tests report scores that are intended to permit comparisons or rankings and use metrics like percentiles and stanines. Criterion-referenced tests report scores in terms of levels of competence or achievement with respect to a performance criterion and use metrics like mastery or percent of an objective achieved. All other differences between norm-referenced and criterion-referenced tests, like the way each is developed and validated, are derived from the need to produce tests that permit the appropriate score interpretation.

Development of Criterion-Referenced Tests

Formulating and generating objectives. One of the basic features of CRTs is their foundation on a clearly-defined set of educational tasks and purposes. CRT objectives can be selected in at least six ways:

1. Expert judgment. Experts assess, on the basis of their knowledge and experience in the field, which educational tasks and purposes are the most important to measure.

2. Consensus judgment. Various groups such as community representatives, curriculum experts, teachers, and/or school administrators decide which educational tasks and purposes they consider to be the most important to measure (Klein, 1972; Wilson, 1973).

3. Curriculum analysis. A team of curriculum experts analyzes a set of curriculum materials in order to identify, and, where necessary, infer the educational tasks and purposes that are the focus of the test (Baker, R.L., 1972).

4. Expert analysis of the subject area to be tested. An in-depth analysis is made of an area--such as mathematics--in order to identify all knowledge and skills that must be acquired if the area is to be learned (Glaser and Nitko, 1971, Nitko, 1973).

5. Theories of learning and instruction. A literature review is conducted and/or consultants called in to formulate series or hierarchies of educational tasks and purposes based upon the results of psychological theory and research (Keesling, J.W., 1975).

6. Empirical studies. Experiments are conducted in order to identify the objectives that are most important, because the skills and knowledge are inherently essential.

No matter how they are derived, educational tasks and purposes are usually called objectives or behavioral objectives. However, it should be noted that these terms have a precise meaning to educators: "An objective is an intent (author's italics) communicated by a statement describing a proposed change in a learner - a statement of what the learner is to be like when he has successfully completed a learning experience" (Mager, 1962).

Developers of CRTs do not always use this definition in its purest sense (Hoepfner, 1975). To them, an objective refers to the content that is supposed to have been learned (e.g., equivalent and nonequivalent sets in sixth-grade math) and sometimes includes the behaviors the student is supposed to exhibit (e.g., naming the first five Presidents of the USA).

Other issues concerning educational tasks and purposes, that is, objectives, relate to the rules needed for writing objectives and how broadly or narrowly they should be stated. Formal rules for generating and stating objectives are needed to ensure the uniformity, manageability, and comprehensiveness of the set of objectives or domain that the CRT measures.*

Still another issue deals with how a domain is organized. The objectives for a single domain can be grouped by grade levels; they can be organized according to major content areas; and/or they can be arranged into a hierarchy according to the complexity of the behaviors involved or the order of instruction.

Formulating and generating items. Once the objectives for the CRT have been chosen, the next step is to construct and/or select test items to measure the objectives. This is one of the most difficult steps in the total developmental process because of the vast number of test items that might

*The set of objectives that a CRT measures is sometimes called a domain or universe of content (Skager, 1975; Cronbach, 1971). However, the term "domain" is used by others to mean the rules for generating test items to measure a specific objective (Hively, et al, 1973).

be constructed for any given objective, even those that are relatively narrowly defined (Klein and Kosecoff, 1973). For example, consider the following objective: "The student can compute the correct product of two single-digit numerals greater than zero where the maximum value of this product does not exceed thirty." The specificity of this objective is quite deceptive since there are fifty-five pairs of numerals that meet this requirement, and at least ten different item types that might be used to assess student performance, as can be seen in Figure 1.

Figure 1

Types of CRT test items using the numerals 3 and 5

The student can compute the correct product of two single-digit numerals greater than zero where the maximum value of this product does not exceed thirty.

a. $\begin{array}{r} 5 \\ \times 3 \end{array}$

b. 5 x 3 =

c. (5)(3) =

d. 5 . 3 =

e. 5 times 3 =

f. The product of 5 and 3 =

g. 5 x _ = 15

h. If x=5 and y=3, what is the value of xy?

i. What numeral multipled by 3 will equal 15?

j. John has 5 apples. Sally has 3 times as many apples as John. How many apples does Sally have?

Further, each of the resulting 550 combinations of pairs and item types could be modified in a variety of ways that might influence whether they were answered correctly. Some of these modifications are:

. vary the sequence of numerals (e.g., 5 then 3 versus 3 then 5)

. use different item formats (e.g., multiple choice versus completion)

. change the mode of presentation (e.g., written versus oral)

. change the mode of response (e.g., written versus oral)

It soon becomes evident that a highly-specific objective could have a potential item pool of well over several thousand items (Hively, 1970, et al, 1973; Bormuth, 1970).

The number of items to construct for each objective is influenced by several factors. Some of these factors are the amount of testing time available and the cost of making an interpretation error, such as saying that a student has achieved mastery when he or she has not. For some objectives, many items are needed in order to obtain a stable estimate of a learner's performance, whereas for other objectives, fewer items will suffice.

A related issue in the construction and generation of CRT items is the degree to which the items should be sampled with respect to their relative difficulty and possible content coverage within an objective. It is a well-known and frequently-used principle of test construction that even slight changes in an item can affect its difficulty. The extent to

which the items within an objective are sampled with respect to difficulty has a direct bearing on the interpretation of the scores obtained. In other words, if only the most difficult items are used, the phrase, "achievement of the objective" has a very different meaning than if the items are sampled over the full range of difficulties.

Another issue concerns a CRT's instructional dependence. The instructional dependence of a CRT refers to the extent to which it is designed for use with a specific educational program (Baker, R.L., 1972; Skager, 1973). CRTs with a greater degree of instructional dependence have objectives and test items that are associated with a particular curriculum or set of educational materials and techniques. CRTs with a smaller degree of instructional dependence, on the other hand, contain objectives and test items that are not necessarily associated with the specific skills or content of an educational program. However, they still may have been developed from several educational programs and consequently, have objectives and items that reflect the bias inherent in these programs. Conversely, CRTs with no instructional dependence are based on a domain of content and behaviors that is independent of any educational program, and therefore, can be to compare several different educational programs.

Consideration of the various issues involved in item generation for CRTs has produced a number of different strategies for generating and constructing items:

1. Panel of experts. A group of measurement and curriculum "experts" decide which items to use based on their knowledge of and experience in the field (Zweig, 1973).

2. Content/process matrix. Basically a variation of the classical test construction technique, this approach involves developing for each objective a matrix of contents and behaviors (or tasks) to be assessed. Items are then systematically sampled within this matrix and perhaps along a third continuum of item difficulty as well (Wilson, 1973).

3. Systematic item generation. Basic "item forms" or specifications are developed for each objective that define the range of item difficulties, all the relevant contents and behaviors, and stimulus and response characteristics of items that can be used to assess the objective (Hively, 1970, et al, 1973; Cronbach, 1971; Skager, 1973; Popham, 1975).

Formulating score interpretation schemes. One of the distinctive features of a CRT is its ability to provide a means for describing what an individual (or group) can do, knows, or feels without having to consider the skills, knowledge, or attitude of others. Consequently, CRT scores are reported and interpreted in terms of the level of performance obtained with respect to the objective(s) or domain on which the CRT is based. This type of score is very different from that used for norm-referenced tests in which scores are reported in terms of the performance of other individuals or groups.

It should be noted that scores on CRT tests need not be limited to just a CRT interpretation. Other score interpretations can also be provided to expand upon the CRT interpretation (Klein, 1970; Cronbach, 1970; Ebel, 1972). An example of one way of combining criterion-and norm-referenced information is: "This school had an average score of 5 out of 10 on the objective (a CRT interpretation) which is one standard deviation below the national average of 7 out of 10 (a norm-referenced interpretation). The idea of using both types of score interpretations is not new and does not reduce the theoretical soundness of the score interpretation (Cronbach, 1970; Klein, 1970, 1971). Combining score interpretations is particularly useful for describing what a student can be expected to be able to do and how exceptional or typical this performance is. Some of the different scores that can be interpreted in a CRT-sense are:

1. "Actual score." The number or percent of items "correct" on a given objective, referring to the number of items actually passed on the test.

2. "True score." An individual's or group's true level of performance on an objective, referring to the portion of the total universe of items for an objective that an individual or group could answer correctly. (That is, if every possible item was tested, this score is the number of items that an individual or group would pass.)

3. "Mastery" of a given objective. This refers to whether an individual or group has achieved a pre-set criterion level of performance. To be legitimate, the criterion level should be meaningful and

preferably empirically justifiable. For example, a criterion level of 7 out of 10 items has meaning if systematic study has shown that those who reach this level can actually do something that others who have not reached this level cannot do, or if baseline data show that the average students achieves this level.

4. Performance time. The time it takes (in class hours or calendar days) for a student to achieve a given performance level.

5. Level readiness. The probability that the student is ready to begin the next level of instruction (this may be based on both the number of items correct and the pattern of answers given to these items).

6. Item difficulty. The percentage of students who "pass" each item; that is, the item's difficulty. (This score is given most often when only one item is tested per objective, for example, National Assessment of Educational Progress.)

7. Total objectives mastered. The number of objectives "passed" or "mastered" by an individual or group.

8. Total individuals who passed. The number of individuals or groups who "passed" or "mastered" each objective.

Validation of CRTs

It is axiomatic that all tests and measures must be field tested before basing decisions upon them. When construction of the objectives and test items is complete, the CRT must be analyzed and validated. This process can involve giving the test to students and studying their responses (response data) or relying upon review by experts (judgmental data).

There is much ambiguity about the procedures appropriate for analyzing CRTs. Nevertheless, there are several dimensions of item and test quality that are considered to be relevant to CRT quality and that have associated with them review procedures, data collection strategies, experimental designs, and statistical indexes.

Establishing item quality. There are several commonly considered dimensions of item quality:

1. Item-objective congruence. A test item is considered "good" if it measures or is congruent with the objective that it is supposed to assess. Item-objective congruence can be established by using judgmental data. Typically, content experts are given a variety of objectives and the items used to measure them, and are asked to assign the items to their appropriate objective, or to comment on the appropriateness of the item-objective relationship.

2. Equivalence (internal consistency within objectives). An item is considered "good" if it "behaves" like other items measuring the same objective. The concept is similar to item-objective congruence, but its proper use depends on response data. Equivalence is usually measured by computing the biserial correlation between the score on an item and the total score on all items measuring that objective.

3. Stability (over time). An item is considered "good" if examinee performance is consistent from one test period to the next in the absence of any special intervention (e.g., instruction is an intervention that can change examinee performance). Stability involves response data and can be measured by using a phi coefficient that correlates scores on the item from two different occasions.

4. Sensitivity to instruction. An item can be considered "good" if it is sensitive to instruction; that is, if the item is able to discriminate between those who have and those who have not benefited from instruction. This measure of item quality is usually computed for CRTs that are linked to particular educational programs and requires response data. Typically, examinees are tested before and after an educational program. Items that many examinees fail before instruction, but pass after instruction, are considered to be sensitive to the instruction.

5. Cultural/sex bias. An item is considered "good" if there are no systematic differences in performance across different cultural groups or sexes. Bias can be assessed using either judgmental or response data. If the former are used, representatives of different cultural groups, members of each sex, and/or linguists examine test items to determine whether vocabulary or content are foreign or could be misinterpreted. If response data are used to assess bias, they are analyzed (typically using ANOVA or regression for item-cultural/sex interactions).

Establishing test quality. There are six dimensions commonly used to express the quality of a CRT:

1. Test-objective congruence. Similar to item-objective congruence, test-objective congruence assesses the extent to which the total test or subtest measures the relevant objective. Test-objective congruence is usually determined by using judgmental data.

2. Equivalence (internal consistency). Test equivalence measures the homogeneity of test items for an objective, that is, how coherently the test items assess the particular objective. This can be measured by using split-half correlation, Kuder-Richardson formulas, or coefficient alpha.

3a. Stability (test-retest or alternate forms). A test is stable to the extent that examinee responses are consistent from one test period to another or across alternate forms of a test in the absence of any intervention.

3b. Stability (number of items per objective and number of objectives per domain). There are two levels at which this type of stability for a CRT can be estimated. At the first level, a determination is made of the number of items that should be tested in order to obtain a stable score on an objective. For this type of stability, the assumption is made that for each objective there is a pool or population of items with mixed difficulties that deals with the objective, and that for any given test a sample of those items is selected. At the second level, a determination is made of the number of objectives that should be tested in order to obtain a stable estimate of performance on the domain. For this type of stability, the assumption is made that a single score is needed that describes an individual's performance on the domain or set of objectives. Stability can be estimated with response data using correlation techniques and/or Bayesian models (Novick and Lewis, 1974).

4. Sensitivity to instruction. Sensitivity to instruction refers to a test's ability to discriminate between those who have and those who have not benefited from instruction. This type of measure of test quality is usually obtained for CRTs that are linked to a specific educational program. It can be measured using response data by comparing test performance before and after instruction or by comparing scores of those who have and those who have not received instruction.

5. Cultural/sex bias. Test bias refers to the existence of systematic differences in test performance across cultural/sex groups. This can be measured by ANOVA or regression techniques using response data or by expert review using judgmental data.

6. Criterion validity. Criterion validity establishes the meaningfulness of the criterion in terms of which CRT scores are interpreted. Establishing criterion validity is either a one-step or a two-step process.

   Step 1: The first step involves assessing the meaningfulness of the domain: that objectives have been selected and organized to be in themselves educationally significant, and that test items have been systematically generated to cover the objectives. Step 1 criterion validity is usually established by having experts review the objectives and test items to determine the extent to which they were developed in conformance with pre-specified procedures, and to which they cover the domain in a comprehensive and meaningful manner.

   Step 1 must be completed for all CRTs, and, in some cases, is sufficient for establishing criterion validity. One example of a CRT that only requires Step 1 criterion validity is a CRT that is based on objectives that are

narrowly-defined and "operationally" stated in such detail that generating test items only requires transposing the objectives into question form. CRT score interpretations for objectives with these characteristics are meaningful because the objectives describe skills that can be measured directly by test items. A second case is when the CRT's objectives are linked to a curriculum and its scores are intended for and interpreted by teachers and curriculum experts. CRT score interpretations in terms of these types of objectives are meaningful because the skills and knowledge being measured are those taught in classrooms using a specific curriculum. A third case in which Step 1 validity is sufficient is when comparative data are provided, or when the CRT score interpretation is supplemented by a normative interpretation, e.g., the class correctly answered an average of 7 out of 10 items, whereas in the district the average class achieved 5 out of 10.

Step 2: In Step 2, criterion validity is established through empirical means, and involves determining whether examinees who perform well on the test have really achieved the educational objective. Step 2 criterion validity can be measured by comparing scores obtained on a CRT by individuals who, in advance of taking the CRT and using <u>independent</u> criteria, were judged to possess or not possess the skills that the objective

is intended to measure. To the extent that the CRT discriminates between these two groups of individuals, the CRT has criterion validity.*

By establishing Step 2 criterion validity, the relationship between test items and the objectives they are supposed to measure is confirmed. Step 2 criterion validity permits assertions about mastery of the individual objectives that comprise a domain and about more complex behaviors whose component parts are defined by the domain. For example, if a reading test has Step 2 criterion validity, then it becomes possible to make statements about mastery of objectives, like: "John Doe can identify the title sentence in a paragraph," and "John Doe can understand main ideas in a reading passage," as well as statements about mastery of a domain, like: "John Doe can read well enough to comprehend daily newspapers or best-selling novels."

Step 2 criterion validity is particularly useful when objectives are not narrowly defined, only a CRT interpretation is provided, and it may be difficult to assume that achievement of the items necessarily reflects achievement of the larger objective or domain.

*Step 2 criterion validity is similar to construct validity, but an objective or a domain, rather than a psychological state, is the construct.

Establishing classical reliability and validity. There has been considerable debate over the appropriateness of "classical" indexes of reliability and validity to criterion-referenced tests. Some psychometricians have argued that since CRT items are selected to measure achievement of specific educational objectives and not to discriminate between students, scores on CRTs can lack variation. This could arise in the following situation: Before instruction, none of the students have mastered the objectives, and they might all receive a score of zero on the criterion-referenced pretest, whereas after instruction, they might all receive very high scores on the criterion-referenced posttest. A lack of variation in student scores, it is claimed, would cause the traditional indexes of reliability and validity (that are based on variance) to be inappropriate (Popham and Husek, 1969).

Others have argued that when CRTs are administered to a heterogeneous sample representing differing degrees of competence and receiving differing instruction on the objective, there will be sufficient variation in test performance to apply the classical statistical formulas (Klein, 1970; Harris, 1973). This latter stance is becoming the accepted view, and it is now held that the classical indexes (e.g., stability, equivalence) can be estimated for CRTs using a heterogeneous population.

CRT's Theoretical Appropriateness for Evaluation Purposes

Relying on the preceding theoretical discussion of the development and validation of CRTs, it is possible to ask:

Based on theoretical considerations alone, are CRTs appropriate to measure achievement for large-scale, effectiveness evaluations?

The answer to this question is yes. An effectiveness evaluation requires instruments that are reliable and valid and provide meaningful scores that can be used to make decisions about educational policy. In theory, there is an orderly set of developmental and validation procedures which, if followed properly, produce CRTs that are based on well-defined sets of objectives and that can provide meaningful and useful score interpretations. Thus, from a theoretical perspective, CRTs are appropriate and desirable for measuring achievement in effectiveness evaluations. However, there are important caveats attached to this conclusion.

- First, there are persons who simply reject the notion of criterion-referenced testing, and with it, the meaningfulness of any CRT score interpretations. If an evaluation is being commissioned by individuals who share this view, then CRTs should not be used since the resulting information, although theoretically sound, is likely to be ignored.

- Second, as is the case with norm-referenced tests, not all CRTs provide the same type of score interpretation. Some CRTs report and interpret scores in terms of the number of items passed per objective, and many educators and policymakers find this type of score interpretation by itself to

be inadequate for most effectiveness evaluation purposes. However, rejection of this type of score interpretation is not equivalent to rejection of the notion of CRTs since there is no reason why CRT scores cannot be supplemented by comparative data.

Review of Currently Available CRTs

In this section, currently available CRTs are reviewed to determine if they are technically sound, and if they have been designed so that they can be easily used for a large-scale effectiveness evaluation. To do this, a list of review criteria were generated and copies of currently available CRTs were obtained from publishers. The CRTs were evaluated using the review criteria. Based on the results of the review, the practical appropriateness of CRTs for evaluation purposes was discussed.

Generating Review Criteria

To structure the review of available CRTs, a set of criteria were generated. The criteria reflect the characteristics generally accepted as being necessary and appropriate for a large-scale effectiveness evaluation. In order to obtain the criteria, several sources were consulted, including a review of the literature, requests for proposals issued by state and federal agencies involving large-scale evaluations, and criteria already-developed and used for reviewing achievement tests. The final set

of criteria were critiqued and approved by senior researchers and administrators on a major evaluation study.

Obtaining CRTs

A list of publishers of educational tests was compiled using test review books (Buros, 1965, 1972; Hoepfner et al., 1970, 1971, 1974), personal contacts, and library sources (Klein and Kosecoff, 1973). It should be noted that publishers on the list were not necessarily known as marketers of CRTs because it was not always possible to predict in advance who published CRTs and who did not, and because it was considered important to include as many publishers as possible in the review.

A letter was sent to each publisher that requested the following information about any criterion-referenced math or reading tests that they might have available.

1. Detailed descriptions of the test battery at each available grade level (e.g., # objectives, # items, subject matter covered...)
2. Sample tests for reading and math at each available grade level
3. Lists of objectives or domains for reading and math at each available grade level
4. Directions for administering and scoring reading and math tests at each available grade level

5. All technical manuals, field test reports, expert reviews, or test analysis information
6. Information about special features like scoring services or cassette-recorded directions
7. Cost information
8. Name and title of person to be contacted for additional information

When publishers' responses were received, they were sorted into three piles: a "totally irrelevant" pile (e.g., tests purporting to measure science, math, handwriting, and aptitude for medical school); a "possibly interesting, but lacking sufficient information for review" pile (e.g., brochures without copies of tests or test manuals; tests of verbal ability, but not reading; responses from individual researchers who had tests that were not ready for publication); and a "potential CRTs" pile (e.g., any publisher who claimed to have a CRT in reading and/or math and who provided, at the minimum, copies of the test(s) and test manuals). Only the 28 CRTs in the third pile were reviewed.

Each CRT was independently reviewed twice using the set of criteria generated for this purpose and discrepancies were resolved by the two reviewers. Any remaining questions, that is, those usually resulting from unclear or insufficient information from the publishers, were followed-up with a phone call to the publisher.

Explaining Review Criteria

There were nineteen criteria against which CRTs were reviewed. (A copy of the forms used by reviewers can be found in Appendix A.) For this review, reading and language arts were considered to be one or math subject area and mathematics a second subject area. All subtests or tests of individual objectives at the same level were grouped together and considered as a single reading or math test. In addition, the criteria were especially designed in order to permit cross-grade level and longitudinal comparisons that typify large-scale evaluations.*

1. Coverage of specific skills. A test must (in the reviewer's opinion) cover skills in reading (language arts) and/or mathematics. Examples of basic skills are reading comprehension, spelling, arithmetic, and telling time as compared to tangential skills like using the library or computation with a slide rule.

2. Grade-level coverage. Forms of the test must be available for grades 1 through 9. (This criterion makes possible comparisons across grade levels as well as longitudinal comparisons).

3. Overlap of objectives across grade levels. In the reviewer's opinion, some or all of the test's objectives must be measured at each grade level in order to make comparisons across grade levels or over time in terms of common educational objectives

*This investigation focused on CRTs that were developed for grades 1-9 since most currently available CRTs have been developed for those grades.

or skills. For this criterion, objectives or test items at different grade levels need not be worded identically. For example, a test item at the second-grade level might have a student read a sentence and select from a series of four pictures, the one that best depicts the sentence; while a parallel but more complex test item at the ninth-grade level might have a student read a paragraph, and select one out of four sentences that best summarizes the paragraph. For this review, the test need not provide a formal means of identifying those test items or objectives that are measured at different grade levels.

4. Number of test forms per grade level. Due to constraints related to test administration and the time available for testing, there should be a limited number of test forms at each grade level. Just one test per grade level is preferred in order to avoid problems with reliability that can arise when several test forms are combined.

5. Complete directions for test administration. A test should provide (in the opinion of the reviewers) thorough and clear instructions for both the examiner and examinee. Directions concerning distributing tests, demonstrating sample questions, and test administration should be provided in a detailed and easy-to-read form.

6. Special equipment needed for test administration. Test administration should not involve any special equipment (like cassettes or visual aids) aside from pencils and scratch paper.

7. Time for testing. A test (reading or math) should be designed to be completed within a given class period. This usually involves no more than a maximum of 40-60 minutes.

8. Group testing. A test must be designed for group administration.

9. Item-objective match. Each test item should be coded to an objective (or the educational tasks and purposes the test claims to measure).

10. Objective coverage. There should be (in the opinion of the reviewers) a sufficient number of items to adequately measure each objective. The number of items per objective should vary as a function of how broadly or narrowly an objective is stated and its level of difficulty.

11. Objective/subjective scoring. A test must use an objective scoring procedure.

12. Machine scoring options. The test must be available in or adaptable to a machine-scoring.

13. Score interpretation scheme. A test must employ a criterion-referenced score interpretation scheme. Tests using CRT interpretations in addition to other types of score interpretation schemes were also acceptable for this criterion.

14. Reusable materials. Due to monetary constraints, it is preferable that test booklets and test manuals be reusable.

15. Curriculum dependence. A test should not be based on the objectives of any particular curriculum or educational program.

16. Costs of tests per pupil. The costs of testing pupils must be affordable for a large-scale study.

17. Formal field test. A test should provide documentation of field test activities. It is preferable that the field test participants be nationally and geographically representative, be a probability sample, and include sufficient numbers of minority persons to estimate bias.

18. Information on item quality. Information should be provided, based either on judgmental or response data, about item stability, sensitivity to instruction, sex/cultural bias, item-objective congruence, and equivalence.

19. Information on test quality. Information should be provided on test quality, based either on judgmental or response data, to include information about internal consistency, test stability, test-objective congruence, sex/cultural bias, sensitivity to instruction, and criterion validity.

Results of the Review

In this section, the results of the twenty-eight tests reviewed for this study are presented. Each individual reading or mathematics test is identified by a numerical code. The codes are necessary because the publishers submitted their materials voluntarily and did not formally consent to a published review. Further, because many of the 28 CRTs were intended for classrooms and not certification evaluation purposes, the review conducted for this investigation tended to make some CRTs look less excellent than they would have if they had been reviewed from another perspective. The names of the publishers whose tests were reviewed can be found in the Appendix.

1. Coverage of specific skills

   Of the twenty-eight tests reviewed, 15 were designed to assess only reading skills, and 13 were designed to assess only mathematics skills. All twenty-eight tests reviewed focused on measuring basic skills in reading and/or mathematics, rather than on tangential skills and thus met the criterion.

2. Grade-level coverage

   Nine tests were available for grades K-9, and thus met the criterion. The remainder varied from CRTs available for grades K-2 to those available for grades K-8.

3. Overlap of objectives across grade levels

Twelve tests appeared to measure the same objectives at all grade levels. Sixteen tests appeared to have some overlapping objectives which were measured at most, but not all, grade levels, depending on "the appropriateness of the objective" and its level of specificity. It should be noted that to make common objectives, test publishers frequently used broadly-stated objectives or skill categories which they then "translated" into tasks and skills of varying complexity for different grade levels.

4. Number of test forms per grade level

Some CRTs had only one test form per grade level and others had as many as 31. Usually those CRTs that offered a limited number of test forms per grade level would include several objectives on a single test form, while those featuring more tests forms per grade-level would assess one or only a few objectives per form. Three tests did not set limits on the number of tests that could be created from their bank of objectives and items.

5. Complete directions for test administration

Twenty-seven of the tests met the criterion by providing adequate directions both to the examiner and examinee for test administration. One test provided for review no information about administration.

6. Special equipment needed for test administration

Twenty-six tests required no special equipment for test administration and, therefore, met the criterion. Two tests required the use of tape recorders or cassettes, and one provided no information. It should be pointed out that many of the 26 tests were specifically designed for use with special equipment and consider its omission to be relatively less desirable.

7. Time for testing

Only two tests met this criterion. Most tests (24) left time for testing open, but from their length appeared to the reviewers to take more than one hour of testing time. One CRT had no information about the time needed for testing.

8. Group testing

Twenty-five tests could be administered to groups and, therefore, met the criterion. Two tests were designed for individual administration only, and one did not provide this information.

9. Item-objective match

Twenty-six tests had each item coded to an objective and one CRT did not provide this information.

10. Objective coverage

The items tested for each objective ranged from 1 to 150 across the 28 tests. (It should be noted that the CRT with 150 items per objective was based on a computerized item bank from which tests of any length could be generated.)

11. Objective/subjective scoring

Twenty-seven tests employed an objective scoring technique, meeting this criterion. One test employed a subjective technique, and one other CRT did not provide this information.

12. Machine scoring option

Eighteen tests met the criterion for machine scoring. Nine CRTs were hand-scorable only, and one CRT did not provide this information.

13. Score interpretation scheme

Twenty-seven tests met the criterion by using some type of criterion-referenced score interpretation scheme. Overwhelmingly, the scheme was expressed as an arbitrary mastery/non-mastery score or the number of items correct on a given objective. Of these same 27 tests, 7 also employed norm-referenced interpretations. One test did not describe its score interpretation scheme.

14. Reusable materials

Twenty-four tests were designed so that at least some portion of the materials could be reused. These usually were the test booklets, when separate answer sheets were provided, and the teacher's and examiner's manuals. Three CRTs had no reusable materials, and one did not provide this information.

15. Curriculum dependence

Twenty-two tests appeared to have total independence from a particular curriculum or instructional program. Six other tests also appeared to be rather general and independent, although they claimed to be based in varying degrees on a review of what is currently being taught in today's schools.

16. Cost of tests per pupil

Based on a purchase of tests in reading or math at the third-grade level, costs ranged from about five cents per student to $6.31 per student. One test had to be implemented at the district level and cost $7500.00. Most tests are sold in sets of 30 - 35 test booklets. To compute costs, it was assumed that an individual student counted 1/30 to 1/35 of the total.

17. Formal field test

Eight tests provide documentation concerning field test activities. However, the information provided was remarkably sparse with several exceptions. Those who did conduct field tests usually attempted to get some sort of geographic and national representation. Fifteen tests claim to have been field tested, but provided no supporting documentation and five additional tests provided no information at all about field tests.

18. Information on item quality

Twelve tests reported having conducted item quality studies based on both response data and/or expert review. Of these, attention typically was paid to item-objective congruence, item stability or equivalence, and sensitivity to instruction. Eight tests reported having some type of review but declined to state the kinds or extent of their studies. Eight other systems did not provide any information at all.

19. Information on test quality

Thirteen tests reported having conducted test quality studies based on response data and/or expert review. Of these, internal consistency, stability, test-objective congruence, sensitivity to instruction, and criterion validity (Step 1) were most frequently attended to. Seven other systems claimed to have performed test quality studies, but provided no supporting documentation. Eight additional systems provided no information at all.

Figure 2 summarizes the results of the review for each test.

Figure 2

Summary of Test Reviews

TEST / CRITERIA	1 Coverage of specific skills	2 Grade level coverage	3 Overlap of objectives across grade levels	4 Number of test forms per grade level	5 Compute directions for test administration	6 Special equipment for test administration	7 Time for testing	8 Group testing	9 Item objective match	10 Objective coverage	11 Objective/Subjective coverage	12 Machine scoreable
001	P	F	S	9-19	P	F	--	P	P	3-5	P	F
002	P	F	S	11-31	P	F	--	P	P	3-5	P	F
003	P	F	S	4	P	P	--	P	P	3-4	P	P
004	P	F	S	1	P	P	--	P	P	1	P	P
005	P	P	S	5-7	P	P	F	P	P	1-40	P	P
006	P	P	S	1-2	P	P	P	P	P	1-40	P	P
007	P	P	S	--	P	P	--	P	P	4	P	P
008*	P	F	A	2	✓	P	P	✓	P	✓	P	P
009	P	P	S	2-6	P	P	--	P	P	1-2	P	P
010	P	P	S	2-3	P	P	--	P	P	1-3	P	P
011	P	F	S	2-9	P	P	--	P	P	2	P	P
012	P	F	S	4-5	P	P	--	P	P	2	P	P
013	P	P	A	5	P	P	--	F	P	45-150	P	F
014	P	F	A	14	P	P	--	F	P	7-27	F	F
015	P	P	S	1	P	P	--	P	P	3	P	P
016	P	P	S	1	P	P	--	P	P	3	P	P
017	P	F	A	3	P	P	--	P	P	5	P	P
018	P	F	A	3	P	P	--	P	P	5-10	P	P
019	P	F	S	--	P	P	--	P	P	1-20	P	P
020	P	F	S	--	P	P	--	P	P	1-20	P	P
021	P	F	A	1	P	P	--	P	P	5-10	P	F
022	P	F	S	1	P	P	--	P	P	5-10	P	F
023	P	F	A	1	P	P	--	P	P	2-5	P	P
024	P	F	A	1	P	P	--	P	P	2-5	P	P
025	P	F	A	1	P	P	✓	P	✓	5-20	P	F
026	P	P	A	1	P	P	--	P	F	36	P	P
027	P	F	A	1	P	P	--	P	P	6-36	P	F
028	P	F	A	1	P	P	--	P	P	50	P	F

Key to Figure 2:

P = Pass
F = Fail
✓ = No information
-- = Open
S = Some
A = Always
N = None

*This test is not yet available to the public

Figure 2 (continued)

Summary of Test Reviews

TEST	CRITERIA	13 Score interpretation: Mastery	# items per objective	Other criterion-referenced	Norm-referenced	Other	No information	14 Reusable materials	15 Curriculum dependence	16 Cost Per pupil tried at grade level(s)	17 Formal Field Test: National score	Geographic scope	Minority representativeness	Probability sampling	Information NOT documented	No information
001		•	•					P	N	$3.25-					•	
002		•	•					P	N	4.50					•	
003		•						P	S	$1.57	•	•				
004		•						P	N	$1.50					•	
005				•	•			P	N	$1.78	•	•	•	•		
006				•	•			P	N	$1.78	•	•	•	•		
007		•	•					P	N	--					•	
008		•		•				P	S	✓						•
009				•	•			P	N	$ .82	•	•	•	•		
010				•	•			P	N	$ .68	•	•	•	•		
011		•						P	S	$2.75-					•	
012		•						P	S	3.61					•	
013		•			•			P	N	$ .81	•	•	•	•		
014				•				P	N	$ .95					•	
015		•						P	N	$1.70					•	
016		•						P	N	$1.48					•	
017			•					P	N	$6.31					•	
018			•					P	N	$5.96					•	
019		•	•					F	N	$7500.					•	
020		•	•					F	N	to start					•	
021			•		--			P	N	$ .05						•
022			•					P	N	$ .05						•
023		•			•			P	S	$1.00	•	•				
024		•			•			P	S	$1.00	•	•				
025		•	•					✓	N	$ .31						•
026		•	•					P	N	$ .75						
027		•						F	N	✓					•	
028							•	P	N	✓					•	•

Key to Figure 2:

P = Pass
F = Fail
✓ = No information
-- = Open

S = Some
A = Always
N = None
• = Some discussion of

Figure 2 (continued)

Summary of Test Reviews

TEST	18 Information on Item Quality						19 Information on Test Quality							
CRITERIA	Sensitivity to instruction	Stability	Sex/Cultural bias	Item-Objective congruence	Information-NOT documented	No information	Internal consistency	Stability	Test-objective congruence	Sex/Cultural Bias	Sensitivity to instruction	Criterion validity	Information-NOT documented	No information
001	•			•							•	•		
002	•		•	•							•	•		
003	•	•							•		•	•		
004		•					•		•		•	•		
005		•		•			•		•	•		•		
006		•		•			•		•	•		•		
007				•					•			•		
008						•								•
009	•		•				•	•	•			•		
010		•		•			•	•				•		
011					•						•	•		
012					•						•	•		
013		•	•	•				•	•	•				
014					•								•	
015				•									•	
016				•									•	
017					•								•	
018					•								•	
019						•								•
020						•								•
021						•								•
022						•								•
023					•								•	
024					•								•	
025						•								•
026					•		•	•	•	•		•		
027						•								•
028						•								•

Key to Figure 2:

P = Pass
F = Fail
✓ = No information
-- = Open

S = Some
A = Always
N = None

CRTs Practical Appropriateness for Effectiveness Evaluation Purposes

Relying on the preceding discussion of the characteristics of currently available CRTs, it is possible to ask:

> *Based on practical considerations alone, are CRTs appropriate for large-scale effectiveness evaluations?*

The answer to this question is no. From the review, it is clear that although no CRT met all the criteria, there are several CRTs that are potentially feasible for effectiveness evaluation purposes. However, using one of these tests would involve considerable effort to adjust it for an evaluation situation. Specifically, the review uncovered some practical problems that diminish currently available CRTs' suitability for an effectiveness evaluation. They are:

1. Many learning objectives. Most of the CRTs reviewed had a large number of very specific learning objectives that were associated with very small units of instruction, like one to five class lessons. The reason for the use of many, narrowly-defined objectives can probably be traced to CRTs' original use by teachers as one of their regular instructional aides in individualizing and evaluating instruction. Nevertheless, an effectiveness evaluation of the impact of just one year of instruction at one grade level, using such a CRT, would generate information about an enormous number of objectives, thus complicating the management, analysis, and reporting of data.

2. Numerous test forms. Many currently available CRTs provide at each grade level separate test forms each measuring just one or a few different objectives. For example, of the 28 tests reviewed some had up to 31 separate test forms per grade level. The appearance of many test forms also probably reflects the original intention to use CRTs as classroom aides. In terms of an effectiveness evaluation, the logistics of administering a number of distinct tests complicates information collection activities and increases the chances of making errors as well as the costs of conducting the evaluation.

3. Time required for testing. Most available CRTs take more than an hour of class time. For example, the review found that 23 of the 28 publishers claimed that their tests were untimed and thus left pacing to the discretion of the examiner; however, based on the number of test items, it is clear that that one hour of test time is insufficient. In terms of the schedules of most evaluation studies, one class period of testing is the maximum time that can usually be devoted to CRT.

It should be noted that some of the test publishers, recognizing time constraints, offered CRTs that had just one item per objective. However, this is not a satisfactory solution since reduction in the number of items will almost invariably bring with it a diminution in the test's ability to measure with precision each of the objectives.

4. Matching CRT's objectives to instruction. Using CRTs in effectiveness evaluations that involve more than one educational program means determining relationships between the CRTs' objectives and the programs' so that achievement can be measured in terms of the objectives emphasized in instruction and exemplary programs can be identified. However, obtaining this information is costly and complicated. Teachers can be asked, for example, to rate the CRTs' objective in terms of their relevance to classroom instruction, but teacher ratings can be unreliable. Instructional experts can be asked to analyze textbooks and curriculum guides; however, they cannot know for certain how these materials are being used in the classroom.

Another problem closely associated to that of relating CRT and instructional objectives concerns which objectives to test. Each student or classroom can be tested on just those objectives that are derived from the curriculum being used; or can be tested on a sample of objectives some of which may be relevant to the curriculum, while the others are not. Depending upon the choice, the resulting evaluation information can be limited in its ability to be used in making comparisons or can require considerable manipulations before interpretations can be made.

5. Identifying common objectives. A fifth problem with using CRTs in effectiveness evaluation studies is that the same objectives are not always measured at all grade levels, or, if they are, there is no system for identifying common objectives. Although the skills and content associated with an objective generally become more complex with increasing grade levels, it is necessary in order to make comparisons over time or across grades to identify skills or objectives that are related in terms of a conceptual framework or general content area. For example, in the fourth grade, a punctuation objective might focus on beginning sentences with capital letters and ending them with periods, while in the ninth grade, a punctuation objective might focus on the proper use of semicolons as alternatives to periods. Although both these objectives deal with the same skill area, grammar, unless they were formally referenced to that general skill area, the evaluator is faced with the responsibility of making this instructional-type of decision, one that is ordinarily not part of in his/her area of expertise.

6. Validating CRTs. The procedures used to validate CRTs are not very sophisticated and field test results are not reported in any detail. When compared with the highly-structured field tests conducted for norm-referenced tests, most CRTs are deficient with respect to the sample's size and representativeness, and/or the amount of precision of data presented in technical reports. It must be

noted that test publishers have probably been reluctant to devote time and money to field testing because test theorists have not been able to provide them with an agreed-upon set of procedures for analyzing and reporting field test data. Assigning blame, however, is not the issue since the fact remains that a paucity of data is provided concerning the technical quality of tests and test items.

7. CRT scores. Most CRTs report scores in one of two ways: either as the number of items correctly answered for each objective, or sometimes as mastery or non-mastery scores, where "mastery" means corectly answering an arbitrarily--selected number of items per objective. These types of score interpretation are accepted by theorists as a legitimate way of expressing CRT test scores and they may have meaning for teachers who know their curriculum. However, for effectiveness evaluation purposes, these types of interpretations alone are inadequate because they provide insufficient information for decision making and lose meaning outside the classroom.

8. Financial considerations. A final practical problem with using currently available CRTs for effectiveness evaluation purposes is that most are costly. This probably reflects the effort it takes to define domains and to produce the special feature offered by CRTs like referencing the objectives to various school curriculums and providing many short test forms that can be used efficiently for classroom instruction purposes.

Conclusions

In previous sections, theoretical and practical characteristics of CRTs were examined. In this section, the results of those examinations are synthesized in order to determine the feasibility of using criterion-referenced tests to measure achievement in an effectiveness evaluation.

The Feasibility of Using CRTs in an Effectiveness Evaluation Context

There is no currently available CRT that is feasible for use in large-scale effectiveness evaluations. This conclusion is based on practical, not theoretical, considerations. One major reason for the likely inappropriateness of available CRTs is that many of them have been designed for classroom and not evaluation, purposes, and consequently, are characterized by numerous, narrowly defined objectives, each measured on a separate test form. In the context of an effectiveness evaluation, these CRTs produce unwieldy amounts of information, require too much time for testing, and create logistical problems for test administrators.

A second major practical failing of currently available CRTs is that field tests are either not documented or are performed inadequately. As a result, the reliability and validity of these CRTs is simply not known, and it is inappropriate to provide decision makers with information of unconfirmed quality.

A third major failing of available CRTs is that the score interpretations given are not as meaningful as can be expected. Most are presented as numbers of items passed, without Step 2 criterion validity information or comparative data as supplements. Other practical findings include the costs of CRTs and the absence of mechanisms for tracking the same skills or objectives across grade levels.

A CRT that is feasible to use to measure achievement in an effectiveness evaluation should be based on a limited set of objectives that represent essential competencies and basic skills, be proven reliable and valid, and be able to provide scores that are meaningful and useful.

References

REFERENCES

Alkin, M.C., Kosecoff, J., Fitzgibbon, C., and Seligman, R. Evaluation and Decision Making: The Title VII Experience, CSE Monograph No. 4 Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Baker, R.L. Measurement considerations in instruction product development. Paper presented at Conference on Problems in Objectives Based Measurement, Center for the Study of Evaluation, University of California, 1972.

Bormuth, J.P. On the Theory of achievement test items. Chicago: University of Chicago Press, 1970.

Buros, O.K. (Ed.). The Mental Measurements Yearbook. Highland Park, New Jersey: Bryphon Press, 1965, 1972.

Cronbach, L.J. Essentials of Psychological Testing. (3rd ed.) New York: Harper, 1970.

Cronbach, L.J. Test validation. In L. Thorndike (Ed.), Educational Measurement (2nd ed). Washington, D.C.: American Council on Education, 1971.

Cronbach, L.J. & Suppes, P., Ed. Disciplined Inquiry for Education National Academy of Education: 1969.

Davis, F.B., and Diamond, J.J. The Preparation of Criterion-Referenced Tests, CSE Monograph No. 3 Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Ebel, R.L. Evaluation and educational objectives: Behavioral and otherwise. Paper presented at the Convention of the American Psychological Association, Honolulu, Hawaii, 1972.

Fink, A. and Kosecoff, J. Evaluation Primer. Book in preparation 1976.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. Measurement in Learning and instruction. In R.L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971, pp. 6520670.

Harris, C. Comments on problems of objectives based measurement. Paper presented at Annual AERA meeting, New Orleans, 1973.

Harris, M.L., & Stewart, D.M. Application of classical strategies to criterion-referenced test construction. A paper presented at the annual meeting of the American Educational Research Association. New York, 1971.

Hively, W. Introduction to domain referenced achievement testing. Symposium presentation, AERA, Minnesota, 1970.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project. CSE Monograph Series in Evaluation, Volume 1. Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Hoepfner, R. et al. CSE Elementary School Test Evaluations. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, 1970.

Hoepfner, R. et al. CSE-ECRC Preschool/Kindergarten Test Evaluations Los Angeles: Center for the Study of Evaluation and Early Childhood Research Center, UCLA Graduate School of Education, 1971.

Hoepfner, R. CSE Secondary School Test Evaluation: Grades 7 & 8. Los Angeles: Center for the Study of Evaluation, UCLA Graduate School of Education, 1974.

Hoepfner, R., 1975. A Theological Examination of Criterion-Referenced Measures Based on Elkin's MEAN test Evaluation Scheme; A Photographic Essay pp. 21-109 Life, October.

Hofstadter, R. Anti-Intellectualism in American Life. Vintage Books, 1963.

Keesling, J.W. Identification of Differing Intended Outcomes and their Implications for Evaluation. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975.

Klein, S.P. Evaluating tests in terms of the information they provide. Evaluation Comment, 1970, 2 (2), 1-6/ ED 045-699.

Klein, S.P. Evaluating Tests in Terms of the Information They Provide, Evaluation Comment, 1971 2 (2).

Klein, S.P. An evaluation of New Mexico's educational priorities. Paper presented at Western Psychological Association, Portland, 1972. TM 002 735. (ED number not yet available.)

Klein, S., Fenstermacher, G., and Alkin, M. "The Center's Changing Evaluation Model," Evaluation Comment, 1971 2 (4).

Klein, S.P., & Kosecoff, J.B. Issues and procedures in the development of criterion-referenced tests. ERIC/TM Report 26. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement and Evaluation, 1973.

Mager, R.F. Preparing instructional objectives. San Francisco: Fearon, 1962.

Nitko, A.J. Problems in the development of criterion referenced tests. Paper presented at Annual AERA Meeting, New Orleans, 1973.

Novick, M.R., and Lewis, C. Prescribing Test Length for Criterion-Referenced Measurement. CSE Monograph No. 3 Center for the Study of Evaluation, University of California at Los Angeles, 1974.

Popham, W.J. Educational Evaluation. New Jersey: Prentice-Hall, 1975.

Popham, W.J., & Husek, T.R. Implications of criterion referenced measurement. Journal of Educational Measurement, 1969, 6 (1), 1-9.

Skager, R. Generating criterion referenced tests from objectives based assessment systems: Unsolved problems in test development, assembly and interpretation. Paper presented at Annual AERA Meeting, New Orleans, 1973.

Skager, R. Critical Differentiating Characteristics for Tests of Educational Achievement, Paper presented at the annual meeting of the American Educational Research Association, Washington D.C. 1975.

Wilson, H.A. A humanistic approach to criterion referenced testing. Paper presented at Annual AERA Meeting, New Orleans, 1973.

Wilson, H.A. A judgmental Approach to Criterion-Referenced Testing, CSE Monograph No. 3, Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Zweig, R., & Associates. Personal communication, March 15, 1973.

APPENDIX A

TESTS REVIEWED

Name of System	Publisher
Fountain Valley Teacher Support System-Reading	Richard Zweig, Association, Inc.
Fountain Valley Teacher Support System-Mathematics	Richard Zweif, Association, Inc.
Prescriptive Reading Inventory	CTB/McGraw-Hill
Diagnostic Mathematics Inventory	CTB/McGraw-Hill
Comprehensive Tests of Basic Skills Form S (CTBS/S)-Reading	CTB/McGraw-Hill
Comprehensive Tests of Basic Skills Form S (CTBS/S)-Mathematics	CTB/McGraw-Hill
ORBIT (Objective's-Referenced Bank of Items and Tests)	CTB/McGraw-Hill
Skills Monitoring System-Reading	Hartcourt, Brace, Javanovich, Inc. (not yet available)
1973 Stanford Reading Tests	Hartcourt, Brace, Javanovich, Inc.
1973 Stanford Mathematics Tests	Hartcourt, Brace, Javanovich, Inc.
Individualized Criterion-Referenced Testing-Reading	Educational Development Corporation
Individualized Criterion-Referenced Testing-Mathematics	Educational Development Corporation
Woodstock Reading Mastery Tests Form A	American Guidance Service
Key Math (Diagnostic Arithmetic Test)	American Guidance Service
Mastery: An Evaluation Tool, SOBAR, Reading	Science Research Associates

TESTS REVIEWED

Name of System	Publisher
Mastery: An Evaluation Tool, Mathematics	Science Research Associates
Individual Pupil Monitoring Systems-Reading	Houghton-Mifflin
Individual Pupil Monitoring Systems-Mathematics	Houghton-Mifflin
Comprehensive Achievement Monitoring (CAM) Maintenance Pkg.-Reading	National Evaluation Systems
Comprehensive Achievement Monitoring (CAM) Maintenance Pkg.-Mathematics	National Evaluation Systems
Objectives-Based Test Sets-Reading	Instructional Objectives Exchange
Objectives-Based Test Sets-Mathematics	Instructional Objectives Exchange
Reading-Analysis of Skills	Scholastic Testing Service
Mathematics-Analysis of Skills	Scholastic Testing Service
Tests of Achievement in Basic Skills (TABS)-Reading	Educational and Industrial Testing Service
Tests of Achievement in Basic Skills (TABS)-Mathematics	Educational and Industrial Testing Service
Reading Inventory Probe I	American Testing Company
Mathematics Inventory Tests	American Testing Company